Energy Research & Social Science 45 (2018) 318–330

Contents lists available at ScienceDirect

Energy Research & Social Science

journal homepage: www.elsevier.com/locate/erss

Original research article

Topic modeling method for analyzing social actor discourses on climate change, energy and food security

L.L. Benites-Lazaro[a,b,*], L. Giatti[a], A. Giarolla[b]

[a] School of Public Health, University of São Paulo, Av. Dr. Arnaldo, 715, São Paulo, 01246-904, Brazil
[b] Earth System Science Center of the National Institute for Space Research, Av. Dos Astronautas, 1758, SJC, 12227-010, Brazil

ARTICLE INFO

Keywords:
Topic modeling
Latent Dirichlet allocation (LDA)
Cluster analysis
Discourse analysis
Climate change
Energy
Ethanol
Food security
Machine learning
Natural language processing

ABSTRACT

Debates on climate change, energy and food security concerns have underscored the need for new methods and tools to explore and understand the complexity of these relevant issues. In this study, we used unsupervised probabilistic modeling—latent Dirichlet allocation (LDA)—to examine the changes in social policy debates related to ethanol production in Brazil and its relationship with climate change and food security. We analyze a large amount of data obtained from Brazilian newspapers, government and business documents, and the bulletins of nongovernmental organizations and social movements from 2007 through 2017. The results from the LDA application allowed us to identify key topics, detect novel trends, and follow them through time, in addition to exhibiting the limitations encountered in identifying social actor discourses. To overcome these limitations, we combine LDA and discourse analysis to enable the construction of topics, concepts, and discourses of the actors surrounding the issue of this study. The findings verify the utility of these techniques by emphasizing the themes and discourses of the actors involved and by identifying the determinant positions of the Brazilian actors in the discussions on ethanol production and its competition with food security and the contributions of ethanol as a renewable energy source to mitigate climate change. This finding provides insights for water–energy–food nexus research.

1. Introduction

The relationship between economic growth and energy consumption [1–3] forms one of the most powerful narratives linked to the rise of the Western world and its techno-economic supremacy. This relationship has influenced worldwide energy policies in both developed and developing countries [4,5]. It involves the consumption of energy extracted from fossil fuels, such that most industrial, agricultural, commercial, domestic, and consumer systems are built around a "plentiful supply of black gold" [6]. Nevertheless, this consumption is associated with a rise in global inequalities and the deterioration of the environment [4]. In particular, the oil used for energy generation has been identified as the largest contributor to greenhouse gas (GHG) emissions, which are adversely affecting the global climate [7].

According to the International Energy Agency (IEA) [8], GHG emissions from the energy sector represent roughly two-thirds of all anthropogenic GHG emissions. Moreover, CO_2 emissions from the energy sector have increased over the past century to historically high levels. Thus, transformative action in the energy sector is a necessity to address the challenges of both climate change mitigation and adaptation. This action would, in turn, enhance the importance of more sustainable ways of using various sustainable energy technologies [9,10]. Given the worldwide interest in renewable fuels to reduce GHG emissions, ethanol from sugarcane has been presented as a clean, low-carbon energy source and a positive solution to the climate crisis and fossil-fuel dependence [10–13]. Particularly, in Brazil, both the government and sugarcane ethanol companies have constructed national narratives about ethanol production. These narratives invoke a connection between the environmental benefits of ethanol and its influence on social well-being with respect to the development of modern technologies for its production [14,15].

However, the use of agro-energy has raised fundamental questions pertaining to the impact of ethanol production, especially with regard to the demand for land and water needed for the production process and the required guarantee of a continuous supply of food crops [16–19]. Energy production and food production have engendered increasing controversy over recent years. Specifically, conflicts have arisen between those promoting agro-energy use and national and global interests [16]. In the last few years, the discourse on food security has been framed as both a national and global problem while

* Corresponding author at: School of Public Health, University of São Paulo, Av. Dr. Arnaldo, 715, São Paulo, 01246-904, Brazil.
E-mail address: lbenites@usp.br (L.L. Benites-Lazaro).

https://doi.org/10.1016/j.erss.2018.07.031
Received 28 November 2017; Received in revised form 23 July 2018; Accepted 26 July 2018
Available online 05 August 2018
2214-6296/ © 2018 Elsevier Ltd. All rights reserved.

being recognized as a fundamental human right [20–22]. This discourse upholds food security and energy security as being closely related to complex environmental, social, political, and economic matters. Thus, regulatory frameworks and governing agents are required to identify the interconnections and interdependencies among sectors with the active participation of government agencies, the private sector, and civil society [23–25].

The challenges of achieving integrated decision-making and coherence in policy-making are particularly acute because sectors operate and policies are executed in areas with differing institutional frameworks at different scales and through several actors [24]. This creates the need for empirical methods and techniques for studying and elucidating the complexity of climate change, such as by outlining the key discussions and relationships involved in agro-energy and food security. These discussions can be conveyed to the public through various communication channels that generate a large amount of textual data. Given the immensity of the data, the use of traditional methods would not be feasible. Thus, new computational tools for organizing, pursuing, and understanding this vast amount of information would be required [26].

Therefore, the main question we raise in this study is related to how big data and machine learning can assist with identifying and exploring the predominant discourses and public discussions, concerns, and interests of various actors; the level of importance that these actors assign to certain issues; and how the related discussions have evolved over the years. In this regard, a large amount of textual data would have to be examined to understand the shifts in the problem over the years to identify a solution and motivate the public to take action [27].

This concern arises because the current major challenges facing humanity include access to clean water, energy production, food security, climate change [23], and a growing world population. It is expected that by 2050, the demand for primary energy will almost double to 80%, while that for food production is expected to rise by more than 50% [28,29]. The competition between food and biofuels will become even more intense in the near future, with the global population estimated to reach nine billion by 2050 [17].

The remainder of this paper is structured as follows. In Section 2, we discuss the existing research on machine learning and big data applied in the social sciences. We explain how topic modeling enables researchers to inductively structure large quantities of text and how discourse analysis enables the identification of the context of predominant discourses. In Section 3, we describe our collection of data for this study and the applications of the data analysis. Then, in Section 4 we present the results from the application of the method, and we discuss the limitations of our method. Finally, in Section 5, we offer the conclusions.

2. Literature background

2.1. Data mining and machine learning techniques

Data mining is defined as the process of exploring massive amounts of data (big data) to discover patterns such as the rules of association or temporal sequences. Accordingly, systematic relationships can be detected between variables, subsets of data, correlations, and trends. Data mining employs the techniques of artificial intelligence, machine learning, pattern recognition, statistical data analysis, and data visualization [30,31].

Data mining and machine learning have long been used in many diverse areas, including finance, marketing, advertising, health, genetics, engineering, and gaming [32–35]. In scientific research (the exact sciences and social sciences), data mining refers to the use of information technology to identify factors, and to discover information related to associations (occurrences linked to a single event), sequences (events linked over time), classification (discovering characteristics), agglomeration (detection of different clusters), prognostics, and the hidden relationships in large databases [31,32,36].

Approximately 80% of electronic data is in a text format [37]. This enormous amount of data, covering almost every area of human endeavor, is not only unclassified and unstructured (or semistructured), but it also contains usable, useless, scientific, and trade-specific data. This has created a significant demand for new powerful tools for converting data into useful, task-oriented knowledge [32]. To satisfy the need to explore, organize, search and understand large numbers of textual information tools and techniques that use machine learning algorithms such as data mining, text mining, and natural language processing (NLP) have been developed [32,37].

The main difference between data mining and text mining is that the data are unstructured in the latter. Text mining thus involves converting unstructured text into structured data, the addition of some derived linguistic features and the elimination of others. Its subsequent insertion into a database can be analyzed later. NLP, in turn, covers a vast span of applications, such as machine translation, natural language generation, sentiment analysis, topic segmentation and recognition, automatic summarization, discourse analysis, speech recognition, and textual analysis with machine/deep learning [38,39].

The interdisciplinary nature of data mining research contributes significantly to its success and extensive applications. The power of data mining can be substantially improved by integrating new methods from multiple disciplines. For example, in mining data containing natural language text, it is reasonable to fuse data mining methods with the methods of information retrieval and NLP [31] or data from social media that provide an enormous amount of continuous real-time data.

2.2. Topic modeling: latent Dirichlet allocation (LDA)

Topic modeling comprises a collection of methods and algorithms [40] that facilitates the organization, understanding, researching, and summarization of a large amount of textual information at a scale impossible for human annotation [41]. Topic modeling is based on the classical methods used in NLP, such as n-gram [42,43] and latent semantic analysis (LSA) [44]. In the n-gram model, it is assumed that documents simultaneously belong to several topics and that topic distributions vary over documents. Its main advantage lies in its relative simplicity of implementation, and its scalability is achieved by increasing the size of n (words). In the LSA probabilistic model, it is assumed that the interdependence between words in a document can be explained by the latent topics to which the document belongs [45]. This interdependence can additionally be explained by a matrix containing word counts per paragraph, wherein the rows represent unique words, and the columns represent each paragraph. The matrix is constructed using a mathematical technique called singular value decomposition (SVD), which reduces the number of rows while preserving the structure of the similarity between columns [44].

LSA has evolved to an information retrieval technique called latent probabilistic latent semantic indexation (pLSI) [46]. pLSI is based on the principle that words used in the same contexts tend to have similar meanings, and its advantage lies in its facility to extract the conceptual content of body text by establishing associations between the given terms that occur in similar contexts [46]. In using the method, the obtained results show that the topics extracted from the documents are conceptually similar, even if the results do not share a specific word or words with the initially established criteria [46].

The most widely used algorithm in topic modeling is LDA [30], which is a Bayesian mixed model for discrete data whereby the elements are assumed to be uncorrelated. This generative probabilistic model provides a powerful tool for discovering and exploring the hidden thematic structures in large datasets [26]. The basic idea behind LDA is that documents are represented as random mixtures of latent topics, and each topic is a probability distribution over a vocabulary [30]. The algorithm can associate words with probabilities [47]. An extension of this method is the correlated topic model, where

correlations among topics are allowed [48].

LDA can be used for many types of data to find patterns, e.g., in genetic data, images, and social networks [26], to analyze business sustainability reports, corporate social responsibility communications [13], public policy [49], etc. Accordingly, a certain topic such as climate change can be examined to reveal its various aspects. In this case, the aspects may include the interests of different actors related to this issue; the participation of actors in climate change negotiations; the major causes of resource depletion, energy use, hazardous emissions, and climate change policies; the relationships between energy use and climate change in the context of politics; and the efforts to limit the effects of climate change [50–52]. We can navigate through time to uncover how discussions of these issues have evolved over the years to explore major cultural shifts. In this way, a more thorough understanding of the problem can be obtained [27]. The thematic structure provides a new framework through which this information can be explored and understood [26].

LDA identifies the themes or topics (we use both terms interchangeably) in a document collection and assumes that each document reflects a combination of topics. In this context, a topic is a list of words that occur in statistically meaningful ways [53]; it is a distribution over a fixed vocabulary of terms [41]. When a document collection is analyzed under these assumptions, probabilistic inference algorithms reveal an embedded thematic structure. With this structure, LDA provides a way to quickly summarize, explore, and search massive collections of documents [47]. In LDA, each document can be viewed as a mixture of several topics, where it is assumed that the distribution of the themes has an a priori Dirichlet distribution. The key in LDA is that any piece of text is composed by selecting words from possible bags-of-words, where each bag corresponds to a topic [53].

Fig. 1 depicts a graphical model of LDA; a node (circle) represents a random variable that is labeled according to its function in the generative process. The hidden nodes—topic proportions, topic assignments, and topics—are unshaded. The solid (shaded) nodes denote the observable variables and represent the collection of words in document. The rectangles are in plate notation, which signify a repetition over the enclosed nodes. The N plate indicates the collection of words within documents, the D plate denotes the collection of documents within the collection, and the K plate denotes the total number of latent topics [26,30]. The latent parameter space therefore consists of θ (the topic-by-document distribution), β (the word-by-topic distribution), Z (the topic indicators of each word in the corpus), and the words W. Both α and η are the hyperparameters controlling the prior conjugate Dirichlet distributions for θ and β, respectively, and significantly influence the behavior of the LDA model. The Dirichlet hyperparameters have a smoothing effect on the multinomial parameters [54]. Reducing this smoothing effect in LDA by lowering the values of α and η will result in more decisive topic associations, therefore, θ and β will become sparser

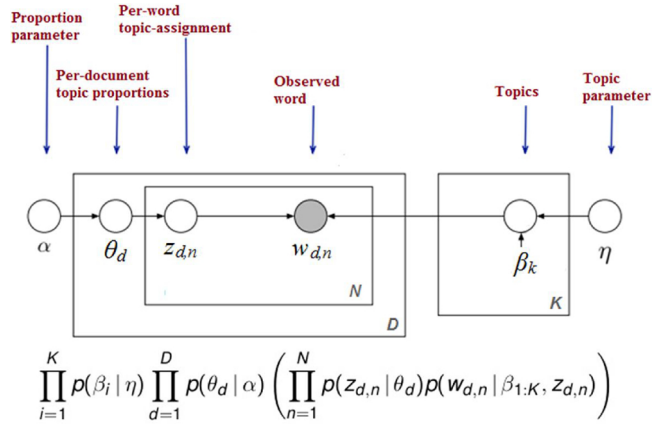

Fig. 1. Latent Dirichlet allocation graphical model [26,30].

[54,55]. The Sparsity of β controlled by η signifies that the model prefers to assign few terms to each topic. The sparsity of θ controlled by α means that the model have a preference to characterize documents by few topics [54].

The inference of these parameters consists of reversing the generative process instead of generating observed words from randomly drawn from β, θ, and Z. We are given the observed words and the work involves estimating these distributions to maximize the log-likelihood of the data. This process of inference is challenging and can be solved using mean-field variational inference [30,56], expectation propagation [57], Gibbs sampling [58], collapsed Gibbs sampling [59], and collapsed variational inference [60]. A detailed explanation of how the LDA method works, a complete derivation of approximate inference using Gibbs sampling, and a method of analysis for the estimation of the hyperparameter α (which is analogous to η) can be found in Heinrich [54].

The LDA model is controlled by three key parameters: the number of topics K and the two hyperparameter vectors α and η. The model demonstrates that thematic analysis can be improved by careful choices of the previous hyperparameters and sampling [61]. That is, learning α and η from the data (given the number of topics K) can be used to increase the model quality, and their estimates reveal the specific properties of the modeled data set [59,62]. The estimate for α is an indicator of how different documents are in terms of their (latent) semantics, and the estimate for η suggests how large the groups of commonly co-occurring words are [59].

In LDA, each document is modeled as a mixture of topics. When the previous θ value is low, the model is forced to a document using fewer topics than when the previous θ value is higher [30,63]. The former is added to the probability of all topics, rendering the distribution of topics more uniform (setting all issues to the same probability). Similarly, topic modeling is performed as a distribution over words. When the previous β (word topic) is high, each topic is forced to balance the assignment of probabilities of each word. Similar to the previous β parameter, it moves the distribution of words in a topic closer to a uniform distribution. The parameter θ controls the number of topics in which the model discretizes the data; increasing this parameter yields more refined themes [30,59,63]. For example, if we only categorize 10 themes in a very large corpus, we cannot discover small topics or micro themes from these selections. As the number of themes increases, the amount of data available to estimate each topic decreases. A large corpus effectively allows the discretization of a large number of topics. The number of topics that best serves the purpose depends on the objectives of the research as well as the size and manner in which the textual data are presented.

2.3. Discourse analysis

Discourse analysis is considered by some researchers as both a theory and a method [64–66]. It consists of a set of techniques for the structured research of texts, and it involves an epistemology that explains how we obtain knowledge of the social world. A set of assumptions concerns the constructive effects of language [67]. Discourse analysis considers how language, both spoken and written, evokes social and cultural perspectives and identities [66]. It then becomes the study of language-in-use and is employed to engender a range of meanings varying from the analysis of linguistic regularities to the normative quality of discussions [68]. The objects of analyses are the ways in which certain problems are represented and in which differences are played out. Social coalitions on specific meanings somehow emerge from the power struggles and entire social context [64].

Depending on the theoretical perspective, some studies have focused on a detailed linguistic analysis of texts, rhetorical devices, constructivist studies that explore diverse ways of reality production, and critical studies that examine the power relations and the formation of discursive knowledge [69]. From a range of perspectives, three

approaches to discourse analysis—discourse theory, critical discourse analysis, and discursive psychology—were described by Jørgensen and Phillips [65]. Each approach has its own distinctive characteristics, namely, philosophical and theoretical premises including particular understandings of discourse and social practices. However, all three approaches have a common aim of conducting critical research, which involves investigating and analyzing the power relations in society [65] from which a critique of such relations can be made to find the possibilities of social change or action for a given problem [27].

Discourse analysis in social sciences is strongly influenced by the work of Michel Foucault [65,68,69], which has played a central role in the development of discourse analysis through both theoretical and empirical research [65]. Specifically, his work on discourses, power, and knowledge follows the social constructivism premise that knowledge is not a reflection of reality but is rather constructed discursively and delimited historically. Power is described by Foucault as both a productive and constraining force closely connected to discourse that is spread across different social practices. Like discourse, power does not belong to particular agents, e.g., individuals or the state, or groups with particular interests; rather, it is spread across different social practices [65].

In environmental discourse analysis, the analyses of leading researchers Hajer [64] and Dryzek [70] have uncovered the power relations within the policy arena. In a study on environmental politics, the three strengths of discourse analysis were highlighted by Hajer and Versteeg [68]: its capacities to reveal the role of language in politics, to reveal the embeddedness of language in practices, and to address questions on how to illuminate mechanisms. Discourse can provide details concerning narratives, beliefs, and needs, and it reveals issues addressed by the text [64,69,71]. The use of discourse can be an important part of wider social and cultural changes, gaining control over the public sphere and communication, where certain societal groups maintain power over others by retaining exclusive access to multitudes of influential public discourses [69,72].

Several studies have displayed how different actors use a particular discourse and impose it onto a discussion [68]. For instance, Clapp and Fuch [73] studied the discursive power of corporations in the U.S. food aid policy by engaging in a public debate over questions of hunger and food security and their links to food aid, showing that grain industries - through persistent public statements- promoted arguments that benefited them. In “Ozone discourses,” Litfin [74] showed how the discursive approach regards international regimes as a localization of power and knowledge providing an arena for contestation among contending discourses with global environmental problems [68]. Kneen [75] in “The Invisible Giant” showed how the Cargill discourse for influencing public opinion was often quoted in terms of economic development, occasionally in humanitarian terms, in an attempt to convince society that the interests of the company are in reality the interests of the country or of humanitarian interest. Furthermore, Benites-Lazaro et al. [15] provided some examples of how, in the last decades, images of nature have evidently become a persistent part of rhetorical strategies of advertising companies. For example, in car advertising, the explicitly or implicitly appropriate growing public concern about the deterioration of the environment is presented as a narrative of harmony between humans, nature, and machines [76,77].

In discourse analysis, there are three ways of approaching text. The traditional, qualitative approach focuses on the way in which the discourse has been delivered. It operates on the premise that the particular manner in which something is said matters. The second approach is quantitative and makes use of statistical tests performed by computer software. The third is a mixed method that takes advantage of both qualitative and quantitative techniques. It assumes that both the number of repetitions of words performed by software and the exploratory lexical patterns play important roles in textual analysis [15,78,79]. In this article, we use the tree approach, which employs a dataset along with two different perspectives. One perspective is qualitative and anchored to discourse analysis; the other is quantitative and involves text classification through software-assisted qualitative data analysis.

3. Material and method

In this study, the source data were documents in text format obtained from Brazilian newspapers (actor MEDIA), reports from non-governmental organizations (actor NGO) and social movements (actor NGO), government websites (actor GOV), and sugarcane ethanol companies represented by the Brazilian Sugarcane Industry Association (actor UNICA) from January 2007 through October 2017. The documents were related to agricultural and agrarian policies, climate change, and biofuel-related issues. The collected documents comprised a large volume of documentary information to be analyzed, for which we made use of several procedures and tools conventionally employed in text mining and machine learning.

3.1. Corpus construction

To scrape a website, we developed scripts using R, an open-source programming language and software environment for statistical computing and graphics. We employed the rvest package to search for HTML code that generates a webpage. To search for websites in JavaScript, we used the PhantomJS package. These packages provide an immediate and easy-to-use solution to search for content generated by free-access web pages such as those of the government, UNICA, and NGO. Access to newspapers that had copyrights was provided by the Observatory on Public Policies for Agriculture at the Rural Federal University of Rio de Janeiro [80]. We used the stringr package to condense the complete transcriptions of 10,200 news articles of approximately 45 MB into a large string in the UTF-8 format and classified the data by month from 2007 to 2017. This task yielded 528 subcorpora (12 documents per year for each of the four actors: GOV, MEDIA, NGO, and UNICA).

3.2. LDA implementation

It was necessary to preprocess the text documents before calculating the thematic model. First, we eliminated empty words, known as stop words such as articles, prepositions, pronouns, adverbs, etc. These are common words that provide no clue about the content of the text. If these stop words remain in the corpus when calculating the model, we would obtain meaningless topics described by these empty words owing to their high probability. Second, a derivation is often performed using stemming algorithms for a specific language, where these algorithms tend to deform the context. It is recommended that a preliminary evaluation be conducted to determine if stemming is appropriate for the desired result. Third, phrases that connote important meanings within a corpus, such as “climate change,” “food security,” are recoded as a single word by replacing the space between the words with an underscore, e.g., “climate change.” A practical method to identify these types of phrases in large corpora involves the use of n-grams.

To model themes, we used Mallet version 2.0.8, a Java-based package for statistical NLP, document classification, clustering, topic modeling, information extraction, and other machine learning applications for text. Mallet uses Gibbs-sampling-based implementations of LDA. We created a directory containing only text files, where each file was considered a document instance. For this study, the corpus was formatted as follows: 10,200 news articles were grouped by month, one month for one document; consequently, during the period of 11 years, we obtained 528 subcorpora, 132 subcorpora for each actor. The thematic analysis could be improved by careful choices of the previous hyperparameters and sampling. Mallet starts with the hyperparameter values of $\alpha = 50/T$ and $\eta = 0.01$ [62] and it requires a definitions of the following input parameters:

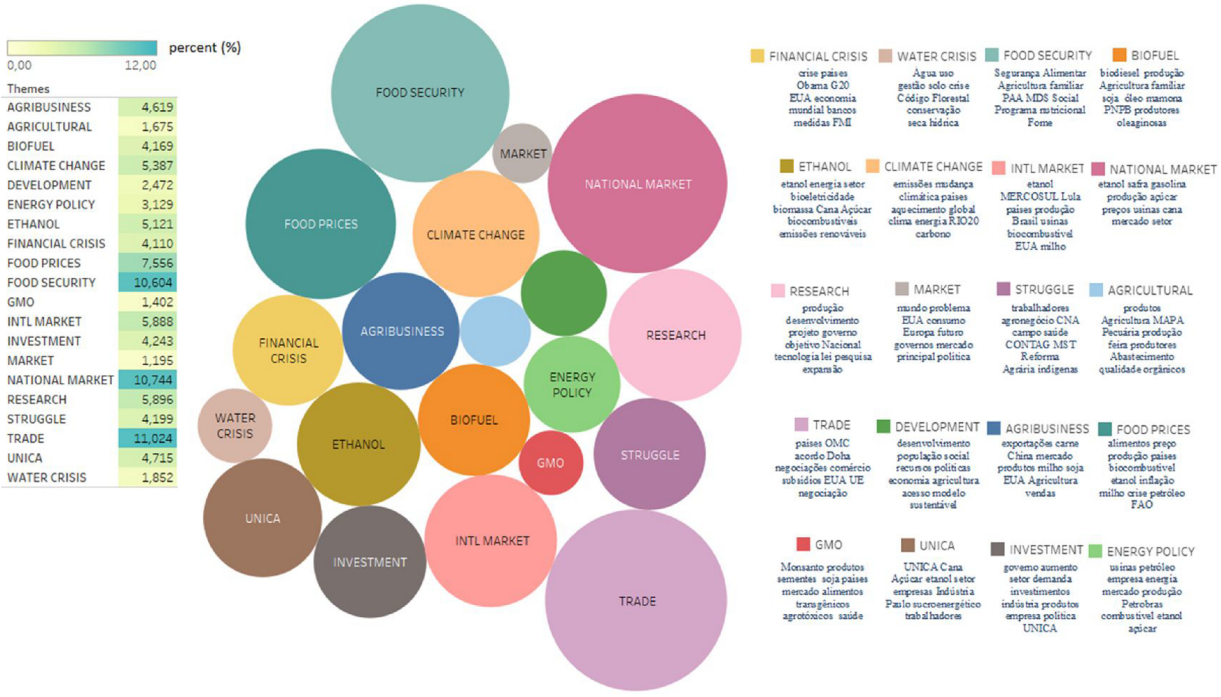

Fig. 2. Themes and top 10 keywords obtained by the application of LDA.

1) Elimination of empty words or stop words: For this purpose, we used a list of stop-words in the Portuguese language.
2) Number of topics: There is no set rule to establish the number of topics. Thus, we followed the recommended number of documents (D) and topics (K), e.g., (D: K) = (1000: 10–20), (10,000: 20–60), and (100,000: 50–200) such that the format of our corpus admitted 20 themes as minimally acceptable.
3) Number of iterations: This number is the number of iterations needed to execute Gibbs sampling. We started with 200, which was the default value for the number of topics. We increased the number of iterations to improve the coherence of the topics. We found 400 to be the optimal number of iterations for our application.
4) Number of thematic words printed: This is the number of words most likely to be printed for each topic following model estimation. The default is to print the top 10 words.
5) Threshold ratio by topic: This parameter was used to prevent printing topics with proportions lower than a threshold value. The default value was 5%.

3.3. Thematic analysis

We used the T-Lab software for the textual analysis of the corpus from the thematic models obtained through LDA using Mallet. This allowed us to find the lexical units (LDA words) and themes (LDA topics) that were “typical” in the subsets of the corpus by extracting the elementary contexts to perform a qualitative analysis of the actors’ discourses. It also facilitated the relationships between the themes and the construction of thematic clusters (LDA topics).

The corpus was loaded using the dictionary-based classification tool of T-Lab that transformed text into a coded corpus set of tables. Finally, this corpus was discretized by T-Lab to yield more than 92,800 words, more than 6,455,000 occurrences, 121,000 contexts, and approximately 65,000 lemmas. We then weighed the elements of the matrix using the term frequency–inverse document frequency (Tf-Idf), where the weight of an element of the matrix is proportional to the number of times that the relevant term appears in each document. Accordingly, the less common terms are up-weighted to reflect their relative importance. We thus obtained approximately 9800 unique words. Processing the text further reduced it to our final subset of approximately 2800 unique words. In the last stage of text preprocessing, T-Lab created different statistical tables to be explored by the different tools of the software.

3.4. Cluster analysis

The objective of the cluster analysis of LDA topics is to find groups of themes that presented two complementary characteristics: maximum internal homogeneity and maximum heterogeneity between groups of topics. These clusters enable the additional analysis of the subsets of the corpus. In addition, with the help of statistical tools such as similarity indices, the chi-squared test, or a p-value, they facilitate the interpretations and inferences pertaining to the meaning of the relationships among the topics and the subcorpora and discourses [81].

The chi-squared test, or goodness-of-fit test, compares the occurrence of a word in a particular corpus (speech of an actor) with its occurrence in other corpora (what is said by the other actors). The similarity indices (cosine index) are used to analyze the co-occurrence of lexical units (LDA words) in the interior of the elementary contexts (subcorpus), where 1 means that two vectors (words with the same similarity) are equal (parallel or complete superposition), and 0 means that they are perpendicular and unrelated to each other. The p-value expresses the statistical significance ($p = 0.05$) [81].

We use cluster analysis to explore the results of LDA modeling for several reasons, for example, the grouping of words and topics adds simplicity and clarity. The most representative themes and words are better highlighted in a 2D map. These graphical representations are visually attractive and the data in the tables are a powerful communication tool.

For the construction of the clusters of the LDA topics, we used the bisecting k-means algorithm (BKM) [82,83] through the T-Lab tool. It is a grouping technique for the type of partitional and hierarchical version of the k-means that uses the minimum sum-of-squares criterion. The BKM algorithm facilitates clustering of our corpus using LDA such as that used in the initial condition. This is equivalent to conducting a supervised clustering process that involves the construction of a contingency table of LDA words multiplied by LDA topics. This is followed

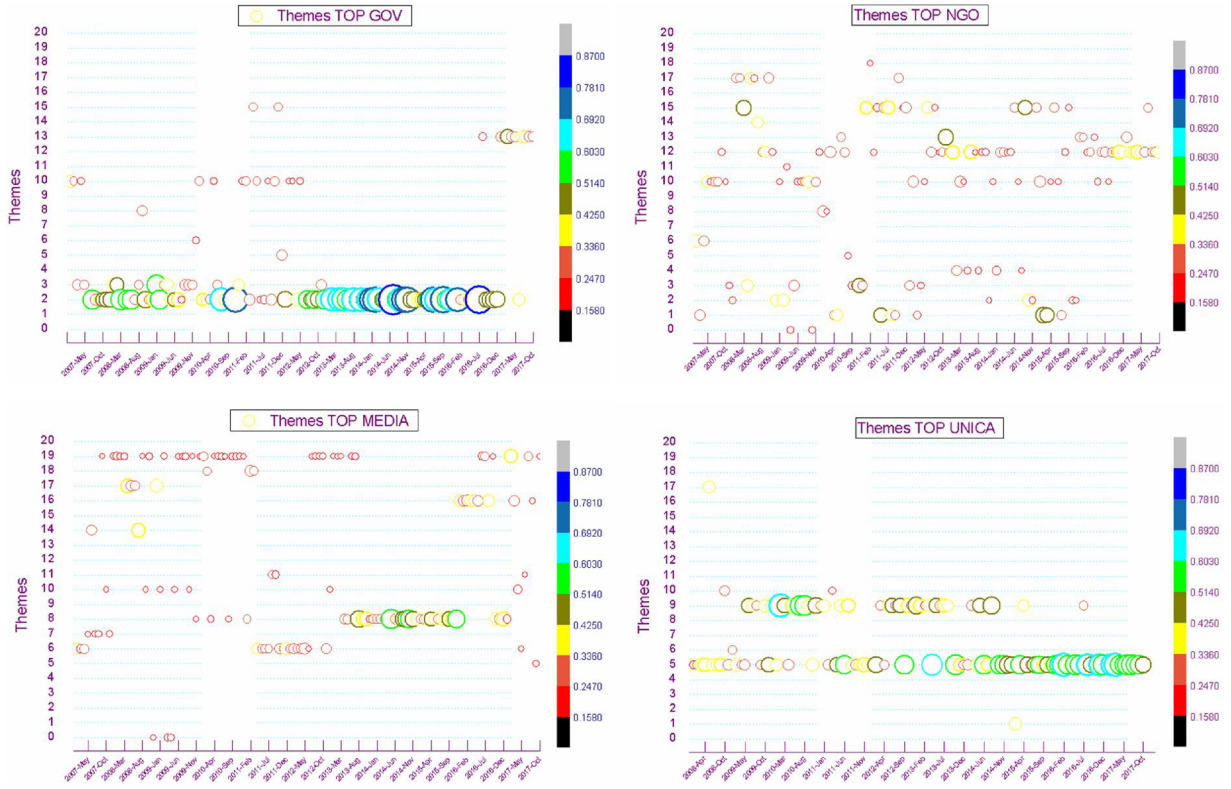

Fig. 3. Timeline of LDA themes by actor from 2007 to 2017.

by a chi-squared test applied to all combinations of LDA words and topics. Finally, a correspondence analysis of the resulting contingency table is conducted.

4. Results and discussion of the application of the method

Fig. 2 shows the 20 main themes with their respective keywords obtained using LDA. These themes are: financial crisis, water crisis, food security, biofuels, GMO-pesticides, ethanol, climate change, international market for ethanol, national market for ethanol, UNICA, research, market access, struggle, agricultural policy, trade, development, agribusiness, food prices, energy policy, and investments. Larger circles indicate themes with more frequent occurrences in the corpus such, as national trade market, and food security.

Fig. 3 and Table 1 present the highest-ranked issues for each actor addressed over time. The actor UNICA highlights “ethanol” and “UNICA.” The latter is a theme related to several activities that this institution performed to promote a positive image of sugarcane companies. The actor GOV emphasizes policies and programs related to food security during all analyzed periods. The actors MEDIA and NGO tended to emphasize various themes. For instance, in the period 2013–2016, MEDIA emphasized the theme of “national markets” related to debates concerning the internal fuel market when the government intervened in the energy market (regarding oil prices) to curb inflation. NGOs assigned importance to almost all themes but focused on certain ones in certain periods, such as themes related to debates on struggles in rural areas, slavery, and the health of peasants.

Fig. 4 depicts the timelines for four themes: climate change, food price, food security, and ethanol. For the topic of climate change, in 2007, the actor NEWS emphasized the release of the Fourth Assessment Report by the Intergovernmental Panel on Climate Change. Furthermore, the first commitment period (2008–2012) of the Kyoto Protocol came into force, followed by the subsequent opening of the carbon market for business and the discussion on the effectiveness of this protocol for mitigating climate change and the participation of countries [52]. The sugarcane sector played an important role in the clean development mechanism of the Kyoto Protocol for the generation of bioenergy projects from sugarcane bagasse. These activities by some companies were related to the sector’s corporate social responsibility activities [84]. In 2009, the greatest attention was given by the actor GOV, when the National Policy on Climate Change was established. This policy Brazil’s voluntary commitment to reduce the projected emissions of greenhouse gases within limits ranging from 36.1% to 38.9% by 2020. The peak in 2011–2012 reflected the United Nations Conference on Sustainable Development (Rio + 20) held in Rio de Janeiro, Brazil, on June 20–22, 2012.

The theme of food prices is a topic for which the four actors participated in discussions. It was the topic of focus for the period 2007–2009, and it accompanied the world food crisis of 2007–2008. For the theme of food security, the actor GOV emphasized food security importance during the entire analyzed period, followed by the actor NGO, whereas ethanol was a theme emphasized by the actor UNICA throughout the period of analysis.

Fig. 5 depicts multidimensional scaling maps obtained using the indexes of similarity (cosine index). These maps enable the interpretation of the relationships that exist among LDA themes and the dimensions that organize the space in which they are represented. The quality of the results is given by the value of the stress function (as its value decreases further below 0.10, the quality of the obtained adjustment becomes greater) [81]; in this case, the stress function is 0.033, which indicates the reliability our results. The themes represented by the blue circles are climate change, food security, and biofuels. The representation indicates that these issues are similar in various respects. For one, they were addressed in 2008, when debates on the controversies related to biofuels and food security gained more attention from the actors. The main discourse from the four actors highlights renewable energy such as ethanol as the contributor to reducing the dependency on fossil fuels, and hence, to mitigating climate change. The discourse also highlights that biofuels are decisive in the fight against global warming and can play important roles in the

Table 1
a) LDA themes by actors. b) LDA themes by years.

a)

Themes	Actors (percentage)			
	GOV	NGO	MEDIA	UNICA
Agribusiness	13.74	10.38	75.72	0.16
Agricultural	47.58	24.67	26.43	1.32
Biofuel	34.87	29.38	34.69	1.06
Climate change	6.85	8.49	82.33	2.33
Development	16.12	32.84	48.96	2.09
Energy policy	1.18	11.79	83.96	3.07
Ethanol	3.31	5.48	46.25	44.96
Financial crisis	0.36	9.16	89.95	0.54
Food prices	3.61	14.75	80.27	1.37
Food security	77.80	13.57	8.56	0.07
GMO pesticides	1.05	36.32	62.11	0.53
Intl market	1.13	6.02	91.85	1.00
Investment	3.48	14.78	80.00	1.74
Market access	1.23	18.52	79.63	0.62
National market	2.68	5.91	84.27	7.14
Research	14.02	23.28	52.69	10.01
Struggle	12.48	70.47	16.34	0.70
Trade	0.54	3.01	95.92	0.54
UNICA	3.76	12.52	36.78	46.95
Water crisis	16.33	32.67	45.82	5.18

b)

Themes	Years distributed (percentage)										
	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Agribusiness	7.35	9.90	5.11	3.04	1.28	8.47	7.19	3.99	15.65	32.91	5.11
Agricultural	1.76	6.17	4.41	2.64	3.08	10.57	8.81	5.29	5.73	29.52	22.03
Biofuel	21.59	17.52	10.09	21.06	4.78	9.38	3.89	3.89	3.01	1.24	3.54
Climate change	36.30	2.19	3.70	2.88	15.21	33.56	0.41	0.55	2.33	1.23	1.64
Development	13.73	7.76	5.07	6.27	15.22	22.99	6.87	5.07	3.88	5.07	8.06
Energy policy	21.23	14.86	16.51	13.92	13.92	4.48	4.01	3.07	3.30	1.89	2.83
Ethanol	13.26	9.94	9.94	13.54	10.23	9.22	4.61	5.76	4.76	10.66	8.07
Financial crisis	8.80	15.80	39.68	10.05	4.13	14.36	2.87	2.15	1.26	0.54	0.36
Food prices	14.26	45.80	5.47	6.45	4.49	7.03	5.27	3.61	3.81	3.52	0.29
Food security	10.51	9.67	4.52	6.26	2.99	10.58	9.88	12.11	14.06	13.92	5.50
GMO pesticides	1.05	1.58	2.63	0.53	1.58	25.79	24.74	11.58	4.74	23.16	2.63
Intl market	47.87	12.53	8.15	4.76	1.63	12.78	5.51	1.75	1.63	2.88	0.50
Investment	17.74	13.74	5.57	7.30	8.52	14.43	8.87	4.70	7.48	6.61	5.04
Market access	17.28	9.88	6.79	8.02	17.90	16.67	8.02	0.00	6.17	6.17	3.09
National market	12.77	9.13	11.33	13.19	3.57	7.62	10.44	11.95	11.40	4.81	3.78
Research	18.40	9.51	7.63	7.26	7.13	12.14	8.51	4.38	7.88	9.14	8.01
Struggle	10.54	9.31	5.80	3.16	4.92	7.56	7.21	4.39	3.34	14.59	29.17
Trade	42.44	36.81	7.63	3.48	0.20	3.28	2.95	0.54	0.80	1.14	0.74
UNICA	14.24	12.05	20.81	17.21	4.54	9.23	6.57	5.32	2.97	2.82	4.23

economic and social developments of the country.

Regarding the competition of ethanol with food crops, the most vigorous defense of sugarcane ethanol came from the Brazilian government, the media, and some NGOs. President Luiz Inacio Lula da Silva at that time (2008) qualified the accusation that ethanol production was contributing to food price inflation as a smoke screen launched by powerful international lobbies. He pointed out that food price inflation results from high oil prices, which affect fertilizer and freight costs; exchange rate changes and speculation in financial markets; falls in world stocks; and increased food consumption in developing countries such as China, India, and Brazil. Furthermore, international NGOs such as Greenpeace, WWF, and Oxfam recognized the virtues of Brazilian ethanol. For instance, Oxfam's position was that Brazilian sugarcane ethanol represented a "special case" among the otherwise harmful ethanol fuels because of its contribution to GHG emission reduction [15]. WWF has actively engaged with the industry using roundtable processes to promote sustainable resource use such as Bonsucro certification [13].

Fig. 6 depicts a correspondence analysis of the themes, which enables the evaluation of the proximity/distance or similarity/dissimilarity between themes and analysis and the identification of the dominant actor discourses. For instance, the results for "food security" highlight several government policies and programs and display the dominant discourse; that is, the case in which food security is a priority of the Brazilian government. The "zero hunger program" was a political strategy based on a set of other programs, such as the federal public food acquisition program (PAA) and the school feeding program (PNAE). Other program aiming to support family-owned farms based on access to credit and insurance is the National Program for the Strengthening of Family Agriculture (PRONAF). With regard to "ethanol," the main discourse highlighted concerns the positive contribution of ethanol to the environment and development. Discourses from the government and companies claiming that ethanol does not compete with the provisioning of food are also highlighted. According to the actor UNICA, ethanol and food security can be simultaneously improved through well-designed bioenergy development. Since sugarcane ethanol production occupies only 1% of the arable land in Brazil, equivalent to 7.8 million hectares, it does not have to compete

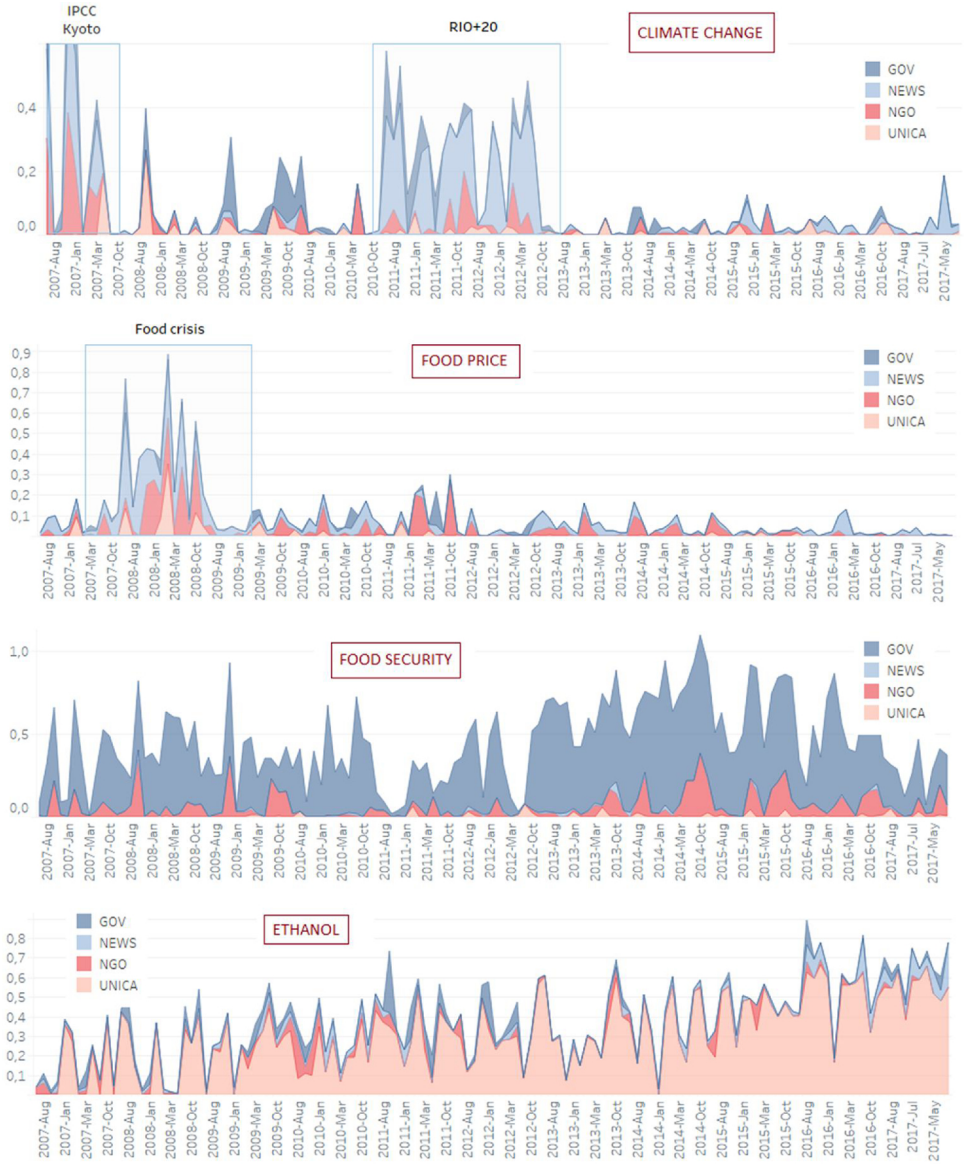

Fig. 4. Linear trends of LDA themes: climate change, food price, food security, and ethanol.

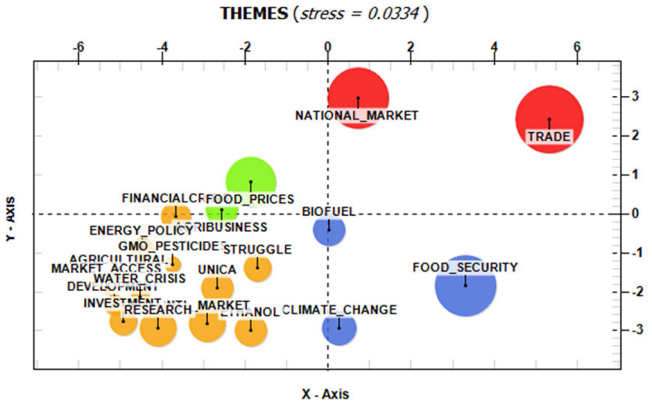

Fig. 5. Multidimensional scaling map of themes.

with crop production for land.

Fig. 7 shows the cluster analysis that was applied for the theme "climate change." This analyses results of the similarities of words classified and translated into semantic relationships. Each pixel represents the similarity of a pair of words, and words with the same similarity are grouped into clusters (subthemes) differentiated by colors. The color of a group shows the clusters of words with significant similarities among themselves and with other words outside the cluster, as denoted in blue for "negotiation and climate agreements," light blue for "effects of climate change," red for "scientific research" on climate change issues, and green for climate change "mitigation actions" with a focus on Brazil's voluntary GHG emission mitigation goals, the National Climate Change Policy Law sanctioned in 2009 (Federal Law no 12187), which established voluntary goals for achieving a reduction between 36.1% and 38.9% of the country's GHG emissions projected to 2020. Although these visual results cannot replace a complete analytical evaluation, they are encouraging and show that the LDA correctly identifies the limits of each subgroup as if discriminated by human reviewers, despite the fact that the algorithm had no knowledge related to the theme.

Through these figures, the results of an unsupervised analysis show that the themes reconstructed by LDA are easily interpretable and coherent, making the automatic learning approach viable in classifying documents for discourse analysis. Moreover, the results indicate that not all issues have the same importance for all actors and that certain

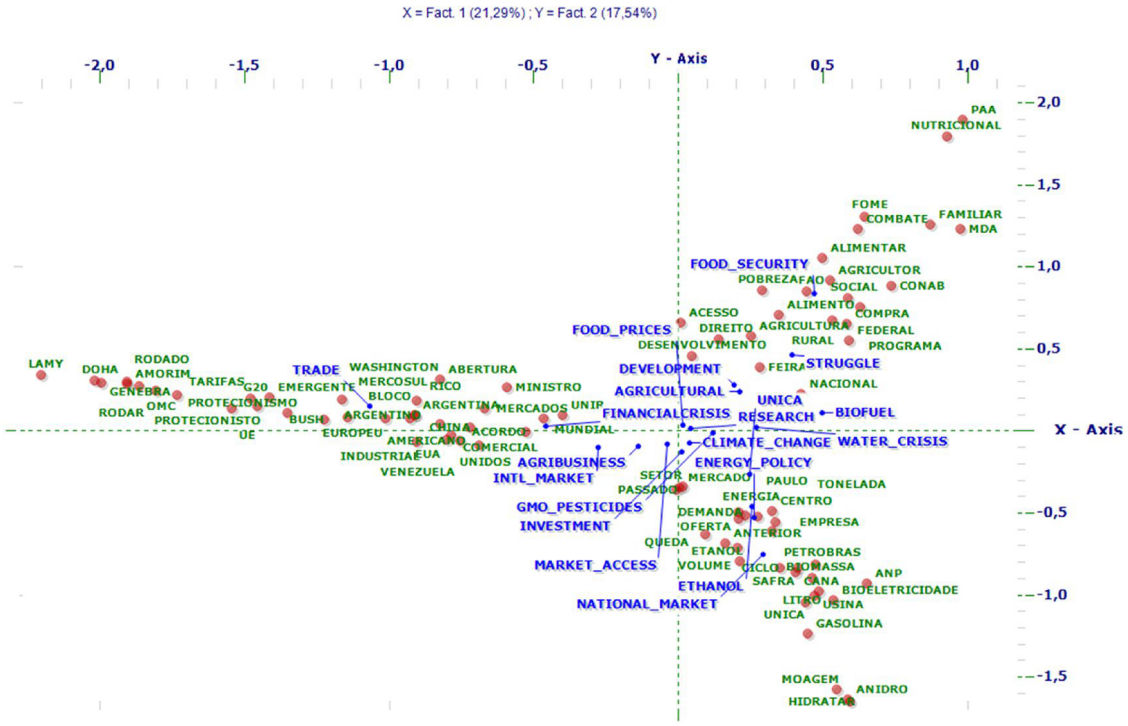

Fig. 6. Correspondence analysis of themes. Themes are labeled in English, and lemmas are in the original Portuguese language of our data. The process of labeling the topics should be treated as an attempt to deduce the meaning of the collection of texts (corpus). Therefore, there is no right or wrong name for the topics, they are only schemes that help to find characteristics or patterns of interest regarding the set of texts that is being studied.

themes do not have the same significance all the time. Accordingly, we consider whether the data matrix is incomplete given the number of issues (Fig. 7). We surmise that this is not necessarily the case because we expect to find themes according to actors with higher densities. Such "agglomeration" occurs in spaces of high dimensionality when the issues are heterogeneous for different actors with different desires and needs. These actors search for blocks of topics that serve their respective purposes, creating varied thematic and discursive sets.

4.1. Discussion on the problems with the method

The results of this study demonstrate that LDA enables the exploration of data, the identification of themes (Fig. 2), the detection of events through time, and the modeling of the evolution of topics over time (Figs. 3 and 4). Despite these techniques being useful tools of inquiry providing quantitative relations that can be measured, topic modeling programs such as LDA know nothing about the meaning of the words in a text. Thus, we combined LDA to generate the themes and a discourse analysis to qualitatively analyze the discourses of the actors through the extraction of the "elemental context" [81]. The "elemental context" fosters the development of insights into how "text and talk" are adapted to their social environments, an act that facilitates this qualitative analysis [15].

The explanation of the results is a fundamentally qualitative endeavor that involves "interpretation, expert knowledge of the specific text material, and calibration of the tool to produce coherent and interpretable topics" [40]. Researchers are the arbiters of meaning from results based on their knowledge of the certain issues [15,85] by considering social, political, historical, and inter-textual contexts [79]. In other words, the qualitative and quantitative aspects of the analysis cannot be neatly separated; instead, they are "integrated as two sides of the same coin" [40].

A benefit of LDA and the segment boundaries is that it outputs the topic distribution associated with each segment [86]. This information is potentially useful to identify themes and the discourses as shown in Figs. 4–7, it is possible to capture the context of the discourses, whereby each actor participates in debates related to the given themes identifying the top interests in certain issues. For example, Fig. 4 shows how the four actors participated in discussions over time as they shared motivations or purposes represented by the latent variables.

One of the challenges in applying LDA, which must be taken into account, is that the results of the model are not deterministic. It has its limitations as any modeling tool. The results are affected by the choices of the input parameters and the incorporated stochastic processes. Some limitations and challenges of the presented method are as follows:

1) Before applying the LDA model, it is necessary to clean the stop words from the corpus. This cleaning can be challenging to undertake; however, the benefit of filtering these words clarifies the generation of topics, improves the accuracy, and reduces the computational load.
2) Regarding the words in the themes, when a theme is general (a few topics), the fact that two words appear as part of a topic does not mean that they appear in the same document. However, as the number of words in the model increases, these words tend to separate to form specific themes. This is shown in the general theme, e.g., in Fig. 7, where the BKM clustering algorithm describes those using "minor" topics or micro themes.
3) There is no set rule to establish a suitable number of themes. Based on our experience, we conclude that the optimal number of LDA topics depends on the aims of the research. In this study, we choose 20 themes that allow the identification of important topics, which are the most representative, meaningful, yet nonredundant, and uninterpretable topics from the data. Early studies shows how researchers used different numbers of topics to achieve their objectives; for example, Evans [87] analyzed public debates from newspapers in the United States with 33 topics, identifying themes such as evolution, climate change, politics, economy, health care, etc, which constitute the debate setting among civil-society actors. Quinn et al. [49] used 42 topics to examine the political text of the

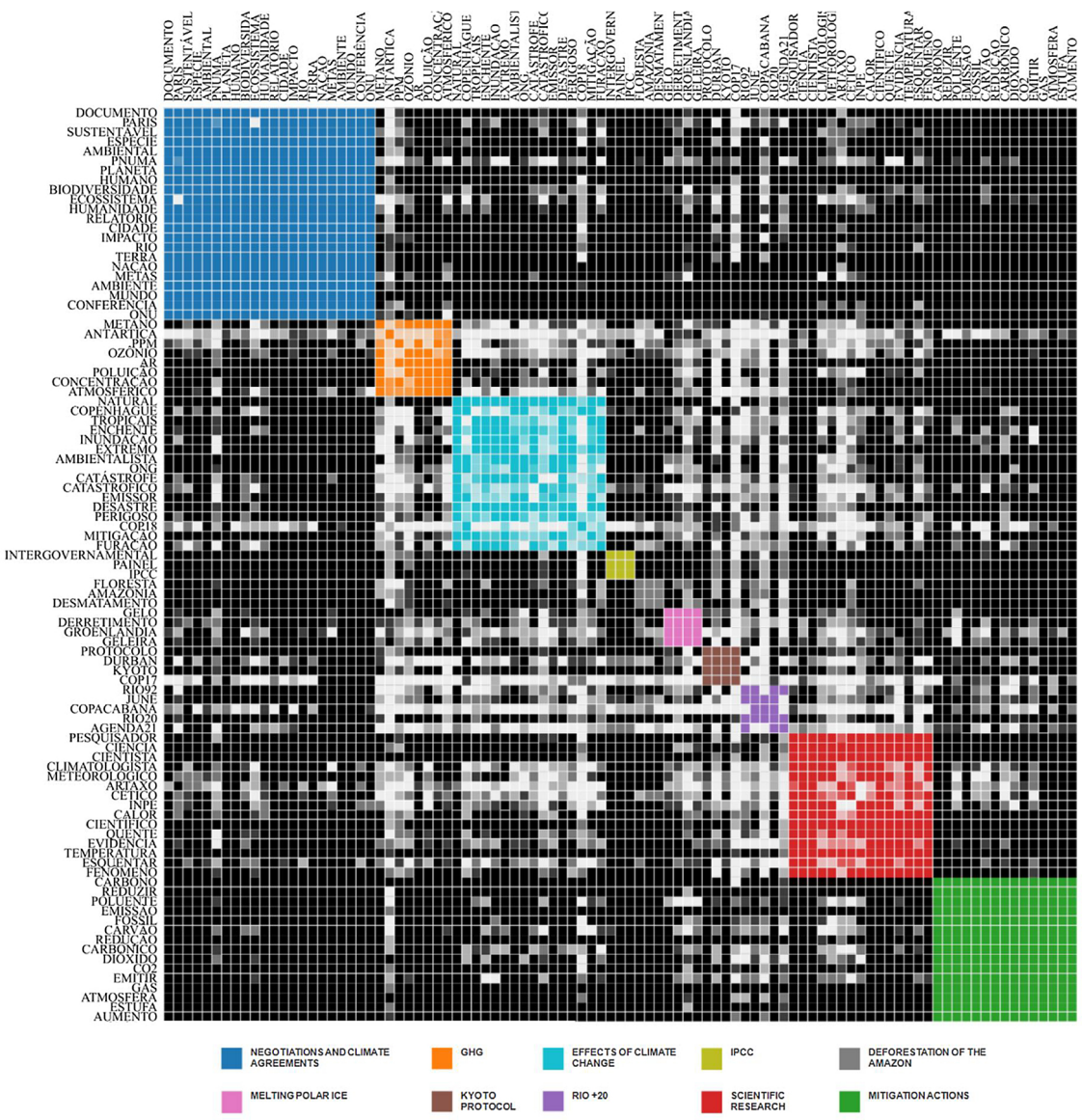

Fig. 7. Cluster analysis of the theme "climate change" and its subthemes.

agenda of the U.S. Senate from 1997 to 2004. Benites-Lazaro et al. [13] used 36 topics to identify companies' commitment to sustainability and business-led governance by analyzing the data from public corporate documents such as sustainability reports and corporate social responsibility communications.

There are also different metrics used to report the process of model selection. One is a measure of the perplexity—the most widely applied, which is used to determine the statistical goodness of fit of a topic model. Another is to run a nonparametric topic model such as a hierarchical Dirichlet process (HDP) in which topics does not need to be defined in advance. Instead, a statistically appropriate number of themes/topics are estimated from the data [88].

Despite, the increasing demand for computational text analysis techniques in the last years, for some researchers in the social sciences that are not well-versed in the use of advanced software that facilitates text processing analysis, it can be challenge and seem daunting. However, for people who are not familiar with environment programs, there is a vast collection of dedicated text analysis packages and text processing from low-level string operations that are easy-to-use. In the Mallet software, for instance, there is an alternative called topic modeling tool (TMT),[1] which is a graphical user interface (GUI) that allows its use without the need to enter a single line of code. The steps required for its use are follows. When executing the TMT, two folders are obtained: output_csv, which contains documents with comma-separated values, and output_html, which contains HTML documents that are visualized by means of a web browser. The output_csv folder has three CSV spreadsheets: DocsinTopics.csv contains the list of topics and their contributions in each document, Topics_Words.csv contains the list of words that define the topics, and TopicsinDocs.csv contains the list of documents and their most important topics. The output_html folder has four HTML documents: all_topics.html, which is the main document; a folder called Docs that contains an HTML page for each document; the Themes folder, which provides an HTML page for each theme; and a Malletgui.css file with instructions to show each HTML page. Another software package that we used in this study is T-Lab; it is one of the most complete, flexible, and easy-to-use packages with linguistic, statistical, and graphical tools for textual analysis [81], it is suitable for novice researchers to use textual analysis techniques more efficiently and fruitfully.

[1] https://code.google.com/p/topic-modeling-tool/.

The traditional approach to discourse analysis received criticism related to its methodological weaknesses such as "the lack of academic rigor in that the analyst's subjective, 'cherry-picked' preconceptions and desired outcome may affect the results of the analysis" [40,89]. It is limited to the qualitative study of small samples of data. However, an alternative involves NLP methods and machine learning techniques for the analysis of large amounts of data. For this purpose, the discursive rhetoric of the different actors is transcribed and converted into textual documents (corpus) that reflect their desires and needs. As a first approximation, the frequency of occurrence in the use of certain words is equivalent to its interest in a specific topic, and the themes are the latent variables that generated the patterns of co-occurrences of opinions (e.g, interests, needs, or associative networks). Furthermore, computer-assisted discourse analysis "is much more eclectic in nature" and "not tied to any particular school of discourse analysis" [89,90]. In fact, it encourages researchers to draw on a variety of analytical tools and frameworks and as much contextual information as may be necessary to obtain the most complete picture [90].

Thus, the combination of the two techniques (LDA and discourse analysis) has allowed us to understand that the discourses of actors on a specific issue were not constant over the years. For example, Fig. 3 shows the topics emphasized by the four actors and the periods in which they were stressed. It is also possible to detect the topics and periods for which the actors' attention coincided. Fig. 4 shows the interest in the theme "food price" when all actors assigned the same importance to it in the same period, 2007–2009. The interest in a specific issue was not similar if actors needed to compete (Fig. 6). Many discourses also achieved an identical goal; therefore, several discourses were "charged" in the same latent variable. Some issues had a similar importance in the discourse of different actors and were therefore said to "belong" to several actors, with each theme defined by a different latent variable.

Nevertheless, in discussions on qualitative and quantitative dichotomy in recent years, a different and concrete approach is emerging in science that is aligned with interdisciplinary approaches. It can link knowledge and expertise to the development of cross-sector dynamics that seek deep interdependencies among sectors [91]. This endeavor faces immense challenges, which are exacerbated by climate change. A key example is the water–energy–food nexus [23] that involves contingency and interdependence among these sectors. Therefore, a need arises for considering a nexus that can break the barriers of sectoral efficiency [91]. Providing clean energy, such as ethanol, can contribute to sustainable development. Nonetheless, ensuring that its production has the best possible effects in other sectors is vital.

This study has implications for the water–energy–food nexus, which is a recent framework that considers interdependencies in a world of deep scarcity. Traditionally, water, energy, and food concerns have been separately handled while neglecting their substantive consequences in terms of trade-offs such as the amount of water needed in all stages of sugarcane and biofuel production [23]. Furthermore, this analysis showed that the question of managing sectors such as those of biofuel/energy and food cannot only focus on companies or governments. Rather, social actors such as the media and NGOs are also relevant to the interactive relationship with society. These considerations are not simply technical issues and must be addressed in terms of social science endeavors [92].

5. Conclusions

In this study, we discussed the utility and limitations of the LDA method in exploring the social policy debates about ethanol production and its competition with food demand obtained from large amounts of data for the last 11 years and the nexus between climate change and energy (ethanol) in Brazil. The results showed that use of only LDA would not allow us to meet our target, as it may have yielded difficulties in understanding the meaning and context of the issue being studied. Thus, we used both LDA and discourse analysis to explore the themes and their contexts and to understand and explain the discourses of the actors as well as the debates and concerns of actors with respect to the issue in a given time period. One of the contributions of this study is the simultaneous application of LDA to four databases of different actors. This allows us to perform a thematic evaluation and capture the similarities and differences that each actor assigns to a theme and its discourses.

The combination of LDA with discourse analysis enabled us to answer the following main questions: "What is the discourse of the various actors on climate change, ethanol, and food security issues in Brazil?"; "What are the topics that have been most often addressed, or the themes to which the greatest importance has been assigned?"; and "Do the actors state the same things all the time?". Machine learning algorithms can be a very powerful tool in the social sciences to provide a different approach to handling complex issues such as climate change and energy (which have deep interdependencies), possible serious scarcity, and the antagonistic discourses of different social actors in a complex arrangement. However, despite theoretical advances and complex mathematical methods in data mining and machine learning, expert knowledge in interpreting and understanding the problem and inferring the context of the discourse remains invaluable.

The context of this study shows that exploring and understanding the nexus of bioenergy and food security in the context of climate change can facilitate the achievement of the 2030 Sustainable Development Goals to end hunger and ensure access to clean energy, as well as the goals of the Paris Agreement under the UN Convention on Climate Change related to bioenergy contribution for reducing GHG emissions. Furthermore, this study underscores the need for integrative governance structures because a clear dichotomy exists between synergy and conflict in the trade-offs between the energy and food sectors. Societal integration in this sense must involve identifying barriers to policy coherence by combining the conventional technical–administrative and sectoral views, thus making it possible to establish shared criteria to guide negotiations by balancing the interests of various social actors.

Thus, future work based on the results of this study will focus on an in-depth analysis of the themes of the climate–water–energy–food nexus in the production of sugarcane ethanol in Brazil. Furthermore, research on LDA topics can be directed to build categorized dictionaries (assignment of varied weights to topics) as indicators of construction. This dictionary can then be applied to other corpora through supervised and fuzzy clustering.

Acknowledgments

The second author acknowledges financial support received from FAPESP - São Paulo Research Foundation (process 2015/50132-6) and CNPq - National Council of Scientific and Technological Development (process 308256/2015-8), Brazil.

References

[1] A. Belke, F. Dobnik, C. Dreger, Energy consumption and economic growth: new insights into the cointegration relationship, Energy Econ. 33 (2011) 782–789.
[2] J.E. Payne, Survey of the international evidence on the causal relationship between energy consumption and growth, J. Econ. Stud. 37 (2010) 53–95.
[3] N. Apergis, J.E. Payne, Energy consumption and growth in South America: evidence from a panel error correction model, Energy Econ. 32 (2010) 1421–1426.
[4] S. Barca, Energy, property, and the industrial revolution narrative, Ecol. Econ. 70 (2011) 1309–1315.
[5] J. Goodman, The "climate dialectic" in energy policy: Germany and India compared, Energy Policy 99 (2016) 184–193.
[6] J. Urry, Consuming the planet to excess, Theory Cult. Soc. 27 (2010) 191–212.
[7] International Energy Agency, CO2 Emissions FromFuel Combustion Highlights, International Energy Agency, Paris, 2015 (Accessed 15 December 2016), http://www.indiaenvironmentportal.org.in/files/file/CO2EmissionsFromFuelCombustionHighlights2015.pdf.
[8] IEA, Energy and Climate Change, IEA, 2015 (Accessed 10 October 2017), https://

www.iea.org/publications/freepublications/publication/WEO2015SpecialReportonEnergyandClimateChange.pdf?.
[9] F. Sengers, R.P.J.M. Raven, A. Van Venrooij, From riches to rags: biofuels, media discourses, and resistance to sustainable energy technologies, Energy Policy 38 (2010) 5013–5027.
[10] J. Goldemberg, S.T. Coelho, P. Guardabassi, The sustainability of ethanol production from sugarcane, Energy Policy 36 (2008) 2086–2097.
[11] E.L. La Rovere, A.S. Pereira, A.F. Simões, Biofuels and sustainable energy development in Brazil, World Dev. 39 (2011) 1026–1036.
[12] G.M. Souza, M.V.R. Ballester, C.H. de Brito Cruz, H. Chum, B. Dale, V.H. Dale, E.C.M. Fernandes, T. Foust, A. Karp, L. Lynd, R. Maciel Filho, A. Milanez, F. Nigro, P. Osseweijer, L.M. Verdade, R.L. Victoria, L. Van der Wielen, The role of bioenergy in a climate-changing world, Environ. Dev. 23 (2017) 57–64.
[13] L.L. Benites-Lazaro, L. Giatti, A. Giarolla, Sustainability and governance of sugarcane ethanol companies in Brazil: topic modeling analysis of CSR reporting, J. Clean. Prod. 197 (2018) 583–591.
[14] E. Malone, N. Hultman, K. Anderson, V. Romero, Stories about ourselves: how national narratives influence the diffusion of large-scale energy technologies, Energy Res. Soc. Sci. 31 (2017) 70–76.
[15] L.L. Benites-Lazaro, N.A. Mello-Théry, M. Lahsen, Business storytelling about energy and climate change: the case of Brazil's ethanol industry, Energy Res. Soc. Sci. 31 (2017) 77–85.
[16] M. Giampietro, K. Mayumi, The Biofuel Delusion: The Fallacy of Large Scale Agro-Biofuels Production, Routledge, 2009.
[17] M.C. Rulli, D. Bellomi, A. Cazzoli, G. De Carolis, P.D. Odorico, The water-land-food nexus of first- generation biofuels, Sci. Rep. 6 (2016) 22521.
[18] R. Rathmann, A. Szklo, R. Schaeffer, Land use competition for production of food and liquid biofuels: an analysis of the arguments in the current debate, Renew. Energy 35 (2010) 14–22.
[19] A. Ajanovic, Biofuels versus food production: does biofuels production increase food prices? Energy 36 (2011) 2070–2076.
[20] R. Sonnino, The new geography of food security: exploring the potential of urban food strategies, Geogr. J. 182 (2016) 190–200.
[21] D. Nally, Governing precarious lives: land grabs, geopolitics, and "food security", Geogr. J. 181 (2015) 340–349.
[22] B.A. Bee, "Si no comemos tortilla, no vivimos:" women, climate change, and food security in central Mexico, Agric. Hum. Values 31 (2014) 607–620.
[23] H. Hoff, Understanding the Nexus. Background Paper for the Bonn 2011 Nexus conference, Stock. Environ. Inst., 2011, pp. 1–52.
[24] N. Weitz, C. Strambo, E. Kemp-Benedict, M. Nilsson, Closing the governance gaps in the water-energy-food nexus: insights from integrative governance, Glob. Environ. Change 45 (2017) 165–173.
[25] S. Larcom, T. van Gevelt, Regulating the water-energy-food nexus: interdependencies, transaction costs and procedural justice, Environ. Sci. Policy 72 (2017) 55–64.
[26] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (2012) 77–84.
[27] A.J. Hoffman, How Culture Shapes the Climate Change Debate, Stanford University Press, 2015.
[28] OECD/FAO, Agricultural Outlook 2013–2022: Highlights, (2013) (Accessed 4 August 2017), https://www.oecd.org/tad/agricultural-policies/OECD-FAO_Outlook_2013-2022.pdf.
[29] IRENA, Renewable Energy in the Water, Energy and Food Nexus, (2015) (Accessed 18 May 2017), http://www.irena.org/documentdownloads/publications/irena_water_energy_food_nexus_2015.pdf.
[30] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[31] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Elsevier-Morgan Kaufmann, 2012.
[32] S. Sumathi, S.N. Sivanandam, Introduction to Data Mining and its Applications, Springer, 2006.
[33] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat. Rev. Genet. 13 (2012) 395–405.
[34] M.J. Shaw, C. Subramaniam, G.W. Tan, M.E. Welge, Knowledge management and data mining for marketing, Decis. Support Syst. 31 (2001) 127–137.
[35] C. Angeli, S.K. Howard, J. Ma, J. Yang, P.A. Kirschner, Data mining in educational technology classroom research: can it make a contribution? Comput. Educ. 113 (2017) 226–242.
[36] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Elsevier, 2017.
[37] V. Gupta, G. Lehal, A survey of text mining techniques and applications, J. Emerg. Technol. Web Intell. 1 (2009) 60–76.
[38] R. Mihalcea, D. Radev, Graph-Based Natural Language Processing and Information Retrieval, Cambridge University Press, 2011.
[39] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
[40] A. Törnberg, P. Törnberg, Muslims in social media discourse: combining topic modeling and critical discourse analysis, Discourse Context Media 13 (2016) 132–142.
[41] D. Blei, J. Lafferty, Topic models, in: A. Srivastava, M. Sahami (Eds.), Text Mining: Classification, Clustering and Applications, CRC Press, 2009, p. 290.
[42] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using EM, Mach. Learn. 39 (2000) 103–134.
[43] T. Dunning, Statistical Identification of Language, Computing Research Laboratory, 1994.
[44] S.T. Dumais, Latent semantic analysis, Annu. Rev. Inf. Sci. Technol. 38 (2004) 188–230.
[45] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (2001) 177–196.
[46] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407.
[47] D. Blei, L. Carin, D. Dunson, Probabilistic topic models, IEEE Signal Process. Mag. 27 (2010) 55–65.
[48] D.M. Blei, J.D. Lafferty, A correlated topic model of science, Ann. Appl. Stat. 1 (2007) 17–35.
[49] K.M. Quinn, B.L. Monroe, M. Colaresi, M.H. Crespin, D.R. Radev, How to analyze political attention with minimal assumptions and costs, Am. J. Pol. Sci. 54 (2010) 209–228, https://doi.org/10.1111/j.1540-5907.2009.00427.x.
[50] R.O. Keohane, D.G. Victor, The regime complex for climate change, Perspect. Polit. 9 (2011) 7–23.
[51] D.L. Levy, P.J. Newell, Business strategy and international environmental governance: toward a Neo-Gramscian synthesis, Glob. Environ. Polit. 2 (2002) 84–101.
[52] L.L.B. Lazaro, A.P. Gremaud, Contribuição para o desenvolvimento sustentável dos projetos de Mecanismo de Desenvolvimento Limpo na América Latina, Organ. Soc. 24 (2017) 53–72.
[53] S. Graham, S. Weingart, I. Milligan, Getting Started with Topic Modeling and MALLET, (2012) (Accessed 7 November 2017), https://uwspace.uwaterloo.ca/handle/10012/11751.
[54] G. Heinrich, Parameter Estimation for Text Analysis, (2009) (Accessed 12 July 2018), http://www.arbylon.net/publications/text-est.pdf.
[55] A. Asuncion, M. Welling, P. Smyth, Y.W. Teh, On smoothing and inference for topic models, Proc. Twenty-Fifth Conf. Uncertain. Artif. Intell. (2009), pp. 27–34.
[56] H. Attias, A variational Bayesian framework for graphical models, Advances in Neural Information Process Systems, MIT Press, 2000, pp. 209–215.
[57] T. Minka, J. Lafferty, Expectation propagation for the generative aspect model, Uncertain. Artif. Intell. (2002), pp. 352–359.
[58] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer, 2004.
[59] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 5228–5235.
[60] Y.W. Teh, D. Newman, M. Welling, A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation, Adv. Neural Inf. Process. Syst. 19 (2007) 1353.
[61] B. Carpenter, B. Baldwin, Text Analysis with LingPipe 4, (2011) (Accessed 20 October 2017), http://alias-i.com/lingpipe-book/lingpipe-book-0.5.pdf.
[62] M. Steyvers, T. Griffiths, Probabilistic topic models, in: T. Landauer, D. McNamara, S. Dennis, W. Kintsch (Eds.), Latent Semantic Analysis. A road to Meaning, Laurence Erlbaum, 2007, pp. 424–440.
[63] H.M. Wallach, D. Mimno, A. Mccallum, Rethinking LDA: why priors matter, Adv. Neural Inf. Process. Syst. 22 (2009) 1973–1981.
[64] M.A. Hajer, The Politics of Environmental Discourse: Ecological Modernization and the Policy Process, Oxford University Press, 1995.
[65] M.W. Jørgensen, L. Phillips, Discourse Analysis as Theory and Method, Sage Publications, London, 2002.
[66] J.P. Gee, An Introduction to Discourse Analysis: Theory and Method, Routledge, 2011.
[67] N. Phillips, C. Hardy, Discourse Analysis: Investigating Processes of Social Construction, Sage Publications, Inc., 2002.
[68] M. Hajer, W. Versteeg, A decade of discourse analysis of environmental politics: achievements, challenges, perspectives, J. Environ. Policy Plann. 7 (2005) 175–184.
[69] N. Fairclough, Analysing Discourse: Textual Analysis for Social Research, Routledge, 2003.
[70] J.S. Dryzek, The Politics of the Earth: Environmental Discourses, second, Oxford University Press, 2005.
[71] T.A. Van Dijk, Discourse and Context: A Sociocognitive Approach, Cambridge University Press, 2008.
[72] A. van Dijk, Society and Discourse: How Social Contexts Influence Text and Talk, Cambridge University Press, 2009.
[73] J. Clapp, D. Fuchs, Agrifood corporations, global governance, and sustainability: a framework for analysis, in: J. Clapp, D. Fuchs (Eds.), Corporate Power in Global Agrifood Governance, The MIT Press, First, 2009, p. 329.
[74] K. Litfin, Ozone Discourses: Science and Politics in Global Environmental Cooperation, Columbia University Press, 1994.
[75] B. Kneen, Invisible Giant: Cargill and its Transnational Strategies, Pluto Press, 2002.
[76] J. Budinsky, S. Bryant, It's not easy being green: the greenwashing of environmental discourses in advertising, Can. J. Commun. 38 (2013) 207–226.
[77] S. Chen, Selling the environment: green marketing discourse in China's automobile advertising, Discourse Context Media 12 (2016) 11–19.
[78] M. Cortini, S. Tria, Triangulating qualitative and quantitative approaches for the analysis of textual materials: an introduction to T-Lab, Soc. Sci. Comput. Rev. 32 (2014) 561–568.
[79] P. Baker, C. Gabrielatos, M. KhosraviNik, M. Krzyżanowski, T. McEnery, R. Wodak, A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press, Discourse Policy Soc. Sci. 19 (2008) 273–306.
[80] OPPA, Presentation, (2017) (Accessed 16 November 2016), http://oppa.net.br/home.php?lang=en.
[81] F. Lancia, T-LAB Tools for Text Analysis, (2017) (Accessed 14 September 2017), http://tlab.it/en/presentation.php.
[82] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, KDD Work. Text Min. (2000), pp. 1–2.
[83] K.L. Judd, On the performance of bisecting K-means and PDDP, Econometrica 53 (1985) 567–585.
[84] L.L. Benites-Lazaro, N.A. Mello-Théry, CSR as a legitimatizing tool in carbon

market: evidence from Latin America's Clean Development Mechanism, J. Clean. Prod. 149 (2017) 218–226.
[85] D. Ketchen, C. Shook, The application of cluster analysis in strategic management research: an analysis and critique, Strateg. Manag. J. 17 (1996) 441–458.
[86] H. Misra, F. Yvon, J.M. Jose, O. Cappe, Text segmentation via topic modeling: an analytical study, Proc. 18th ACM Conf. Inf. Knowl. Manag. (2009), pp. 1553–1556.
[87] M.S. Evans, A computational approach to qualitative analysis in large textual datasets, PLoS One 9 (2014) e87908.
[88] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri, S. Adam, Applying LDA topic modeling in communication research: toward a valid and reliable methodology, Commun. Methods Meas. 12 (2018) 93–118.
[89] P. Baker, E. Levon, Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity, Discourse Commun. 9 (2015) 221–236.
[90] S. Jaworska, A comparative corpus-assisted discourse study of the representations of hosts in promotional tourism discourse, Corpora 11 (2016) 83–111.
[91] P.R. Jacobi, L. Giatti, Nexus for sustainability: searching for a new rationality, Ambient. Soc. 20 (2017) i–ii.
[92] R. Cairns, A. Krzywoszynska, Anatomy of a buzzword: the emergence of "the water-energy-food nexus" in UK natural resource debates, Environ. Sci. Policy 64 (2016) 164–170.